

# A Note on Randomized Element-wise Matrix Sparsification

Abhisek Kundu \*

Petros Drineas †

## Abstract

Given a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , we present a randomized algorithm that sparsifies  $\mathbf{A}$  by retaining some of its elements by sampling them according to a distribution that depends on both the square and the absolute value of the entries. We combine the ideas of [4, 1] and provide an elementary proof of the approximation accuracy of our algorithm following [4] without the truncation step.

## 1 Introduction

Element-wise matrix sparsification was pioneered in [2, 3] and was later improved in [4, 1]. More specifically, the original work of [2, 3] sampled entries from a matrix with probabilities depending on the square of an entry for “large” entries and on the absolute value of an entry for “small” entries. [4] proposed to zero out the small entries and then used sampling with respect to the squares of the remaining entries in order to sparsify the matrix; an elegant proof was possible via a matrix-Bernstein inequality. Very recently, [1] argued that the zeroing out step could be avoided by sampling with respect to the absolute values of the matrix entries. Theorem 1 combines the ideas of [4, 1] to provide an elementary proof that bypasses the zeroing out step. More specifically, we avoid zeroing out the small elements of the input matrix by constructing a sampling probability distribution that depends on both the absolute values *as well as* the squares of the entries of the input matrix.

## 2 Our Result

We present our main algorithm (Algorithm 1) and the related Theorem 1, which is our main quality-of-approximation result for Algorithm 1.

### 2.1 Notation

We use bold capital letters (e.g.,  $\mathbf{X}$ ) to denote matrices and bold lowercase letters (e.g.,  $\mathbf{x}$ ) to denote column vectors. Let  $[n]$  denote the set  $\{1, 2, \dots, n\}$ . We use  $\mathbb{E}(X)$  to denote the expectation of a random variable  $X$ ; when  $\mathbf{X}$  is a random matrix,  $\mathbb{E}(\mathbf{X})$  denotes the element-wise expectation of each entry of  $\mathbf{X}$ . For a matrix  $\mathbf{X} \in \mathbb{R}^{m \times n}$ , the Frobenius norm  $\|\mathbf{X}\|_F$  is defined as  $\|\mathbf{X}\|_F^2 = \sum_{i,j=1}^{m,n} \mathbf{X}_{ij}^2$ , and the spectral norm  $\|\mathbf{X}\|_2$  is defined as  $\|\mathbf{X}\|_2 = \max_{\|\mathbf{y}\|_2=1} \|\mathbf{X}\mathbf{y}\|_2$ . For symmetric matrices  $\mathbf{A}, \mathbf{B}$  we say that  $\mathbf{B} \succeq \mathbf{A}$  if and only if  $\mathbf{B} - \mathbf{A}$  is a positive semi-definite matrix.  $\mathbf{I}_n$  denotes the  $n \times n$  identity matrix and  $\ln x$  denotes the natural logarithm of  $x$ . Finally, we use  $\mathbf{e}_i$  to denote standard basis vectors whose dimensionalities will be clear from the context.

---

\*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, kundua2@rpi.edu.

†Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, drinep@rpi.edu.

## 2.2 Algorithm

Our main algorithm (Algorithm 1) randomly samples (in independent, identically distributed trials)  $s$  elements of a given matrix  $\mathbf{X}$  according to a probability distribution  $\{p_{ij}\}_{i,j=1}^{m,n}$  over the elements of  $\mathbf{X}$ .

---

### Algorithm 1 Element-wise Matrix Sparsification Algorithm

---

- 1: **Input:**  $\mathbf{X} \in \mathbb{R}^{m \times n}$ ,  $\{p_{ij}\}_{i,j=1}^{m,n}$  such that  $p_{ij} \geq 0$  (for all  $i, j$ ) and  $\sum_{i,j=1}^{m,n} p_{ij} = 1$ , integer  $s > 0$ .
  - 2: **For**  $t = 1 \dots s$  (i.i.d. trials with replacement) **randomly sample** pairs of indices  $(i_t, j_t) \in \{1 \dots m\} \times \{1 \dots n\}$  with  $\mathbb{P}[(i_t, j_t) = (i, j)] = p_{ij}$ .
  - 3: **Output:** set of sampled pairs of indices  $\Omega = \{(i_t, j_t), t = 1 \dots s\}$ .
  - 4: **Sampling operator:**  $\mathcal{S}_\Omega : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$  with  $\mathcal{S}_\Omega(\mathbf{X}) = \frac{1}{s} \sum_{t=1}^s \frac{\mathbf{X}_{i_t j_t}}{p_{i_t j_t}} \mathbf{e}_{i_t} \mathbf{e}_{j_t}^T$ .
- 

**Theorem 1** Let  $\mathbf{X} \in \mathbb{R}^{m \times n}$  and let  $\epsilon > 0$  be an accuracy parameter. Let  $\mathcal{S}_\Omega : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$  be the sampling operator of the element-wise sampling algorithm (Algorithm 1) and assume that the sampling probabilities  $\{p_{ij}\}_{i,j=1}^{m,n}$  satisfy

$$p_{ij} \geq \frac{\beta}{2} \left( \frac{\mathbf{X}_{ij}^2}{\|\mathbf{X}\|_F^2} + \frac{|\mathbf{X}_{ij}|}{\sum_{i,j=1}^{m,n} |\mathbf{X}_{ij}|} \right) \quad (1)$$

for all  $i, j$  and some  $\beta \in (0, 1]$ . Then, with probability at least  $1 - \delta$ ,

$$\|\mathcal{S}_\Omega(\mathbf{X}) - \mathbf{X}\|_2 \leq \epsilon,$$

if either (i)  $\epsilon \leq \|\mathbf{X}\|_F$  and  $s \geq \frac{6 \max\{m, n\} \ln((m+n)/\delta)}{\beta \epsilon^2} \|\mathbf{X}\|_F^2$ ,  
or (ii)  $\epsilon > \|\mathbf{X}\|_F$  and  $s \geq \frac{6 \max\{m, n\} \ln((m+n)/\delta)}{\beta \epsilon} \|\mathbf{X}\|_F$ .

We now restate the above bound in terms of the stable rank of the input matrix. Recall that the stable rank is defined as  $\mathbf{sr}(\mathbf{X}) := \|\mathbf{X}\|_F^2 / \|\mathbf{X}\|_2^2$  and is upper bounded by the rank of  $\mathbf{X}$ .

**Corollary 1** Let  $\mathbf{X} \in \mathbb{R}^{m \times n}$ , let  $\epsilon > 0$  be an accuracy parameter such that  $\mathbf{sr}(\mathbf{X}) \geq \epsilon^2$ , and let  $\mathcal{S}_\Omega(\mathbf{X})$  be the sparse sketch of  $\mathbf{X}$  constructed via Algorithm 1 with the  $p_{ij}$ 's satisfying the bounds of eqn. (1). If

$$s \geq \frac{6 \max\{m, n\} \ln((m+n)/\delta)}{\beta \epsilon^2} \mathbf{sr}(\mathbf{X}),$$

then, with probability at least  $1 - \delta$ ,

$$\|\mathbf{X} - \mathcal{S}_\Omega(\mathbf{X})\|_2 \leq \epsilon \|\mathbf{X}\|_2.$$

## 3 Proof of Theorem 1

In this section we provide a proof of Theorem 1 following the lines of [4]. First, we rephrase the non-commutative matrix-valued Bernstein bound theorem of [5] using our notation.

**Theorem 2** [Theorem 3.2 of [5]] Let  $\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_s$  be independent, zero-mean random matrices in  $\mathbb{R}^{m \times n}$ . Suppose  $\max_{t \in [s]} \{ \|\mathbb{E}(\mathbf{M}_t \mathbf{M}_t^T)\|_2, \|\mathbb{E}(\mathbf{M}_t^T \mathbf{M}_t)\|_2 \} \leq \rho^2$  and  $\|\mathbf{M}_t\|_2 \leq \gamma$  for all  $t \in [s]$ . Then, for any  $\epsilon > 0$ ,

$$\left\| \frac{1}{s} \sum_{t=1}^s \mathbf{M}_t \right\|_2 \leq \epsilon$$

holds, subject to a failure probability of at most

$$(m+n) \exp\left(\frac{-s\epsilon^2/2}{\rho^2 + \gamma\epsilon/3}\right).$$

For all  $t \in [s]$  we define the matrix  $\mathbf{M}_t \in \mathbb{R}^{m \times n}$  as follows:

$$\mathbf{M}_t = \frac{\mathbf{X}_{i_t j_t}}{p_{i_t j_t}} \mathbf{e}_{i_t} \mathbf{e}_{j_t}^T - \mathbf{X}. \quad (2)$$

It now follows that

$$\frac{1}{s} \sum_{t=1}^s \mathbf{M}_t = \frac{1}{s} \sum_{t=1}^s \left[ \frac{\mathbf{X}_{i_t j_t}}{p_{i_t j_t}} \mathbf{e}_{i_t} \mathbf{e}_{j_t}^T - \mathbf{X} \right] = S_\Omega(\mathbf{X}) - \mathbf{X}.$$

Let  $\mathbf{0}_{m \times n}$  denote the  $m \times n$  all-zeros matrix and note that  $\mathbf{X} = \sum_{i,j=1}^{m,n} \mathbf{X}_{ij} \mathbf{e}_i \mathbf{e}_j^T$ . The following derivation is immediate (for all  $t \in [s]$ ):

$$\mathbb{E}(\mathbf{M}_t) = \mathbb{E}(S_\Omega(\mathbf{X})) - \mathbf{X} = \sum_{i,j=1}^{m,n} p_{ij} \frac{\mathbf{X}_{ij}}{p_{ij}} \mathbf{e}_i \mathbf{e}_j^T - \mathbf{X} = \mathbf{0}_{m \times n}.$$

The next lemma bounds  $\|\mathbf{M}_t\|_2$  for all  $t \in [s]$ .

**Lemma 1** Using our notation,  $\|\mathbf{M}_t\|_2 \leq \frac{3\sqrt{mn}}{\beta} \|\mathbf{X}\|_F$  for all  $t \in [s]$ .

*Proof:* Notice that sampling according to the element-wise probabilities of eqn. (1) satisfies

$$p_{ij} \geq \frac{\beta}{2} \frac{|\mathbf{X}_{ij}|}{\sum_{i,j=1}^{m,n} |\mathbf{X}_{ij}|}.$$

We can use the above inequality to get

$$\|\mathbf{M}_t\|_2 = \left\| \frac{\mathbf{X}_{i_t j_t}}{p_{i_t j_t}} \mathbf{e}_{i_t} \mathbf{e}_{j_t}^T - \mathbf{X} \right\|_2 \leq \frac{2}{\beta} \sum_{i=1}^m \sum_{j=1}^n |\mathbf{X}_{ij}| + \|\mathbf{X}\|_2 \leq \frac{3\sqrt{mn}}{\beta} \|\mathbf{X}\|_F.$$

In the above we used  $\beta \leq 1$ ,  $\|\mathbf{X}\|_2 \leq \|\mathbf{X}\|_F$ , and (from the Cauchy-Schwarz inequality)

$$\sum_{i,j=1}^{m,n} |\mathbf{X}_{ij}| \leq \sqrt{mn \sum_{i,j=1}^{m,n} \mathbf{X}_{ij}^2} = \sqrt{mn} \|\mathbf{X}\|_F.$$

Thus, we get a new bound for Lemma 2 of [4], bypassing the need for a truncation step.  $\diamond$

Next we bound the spectral norm of the expectation of  $\mathbf{M}_t \mathbf{M}_t^T$ . The spectral norm of the expectation of  $\mathbf{M}_t^T \mathbf{M}_t$  can be bounded using a similar analysis.

**Lemma 2** Using our notation,  $\|\mathbb{E}(\mathbf{M}_t \mathbf{M}_t^T)\|_2 \leq \frac{2n}{\beta} \|\mathbf{X}\|_F^2$  for all  $t \in [s]$ .

*Proof:* Recall that  $\mathbf{X} = \sum_{i,j=1}^{m,n} \mathbf{X}_{ij} \mathbf{e}_i \mathbf{e}_j^T$  and  $\sum_{i,j=1}^{m,n} p_{ij} = 1$  to derive

$$\begin{aligned}
\mathbb{E}[\mathbf{M}_t \mathbf{M}_t^T] &= \mathbb{E} \left[ \left( \frac{\mathbf{X}_{i_t j_t}}{p_{i_t j_t}} \mathbf{e}_{i_t} \mathbf{e}_{j_t}^T - \mathbf{X} \right) \left( \frac{\mathbf{X}_{i_t j_t}}{p_{i_t j_t}} \mathbf{e}_{j_t} \mathbf{e}_{i_t}^T - \mathbf{X}^T \right) \right] \\
&= \sum_{i,j=1}^{m,n} p_{ij} \left( \frac{\mathbf{X}_{ij}}{p_{ij}} \mathbf{e}_i \mathbf{e}_j^T - \mathbf{X} \right) \left( \frac{\mathbf{X}_{ij}}{p_{ij}} \mathbf{e}_j \mathbf{e}_i^T - \mathbf{X}^T \right) \\
&= \sum_{i,j=1}^{m,n} \left( \frac{\mathbf{X}_{ij}^2}{p_{ij}} \mathbf{e}_i \mathbf{e}_j^T \mathbf{e}_j \mathbf{e}_i^T \right) - \left( \sum_{i,j=1}^{m,n} \mathbf{X}_{ij} \mathbf{e}_i \mathbf{e}_j^T \right) \mathbf{X}^T - \mathbf{X} \left( \sum_{i,j=1}^{m,n} \mathbf{X}_{ij} \mathbf{e}_j \mathbf{e}_i^T \right) + \sum_{i,j=1}^{m,n} p_{ij} \mathbf{X} \mathbf{X}^T \\
&= \sum_{i,j=1}^{m,n} \left( \frac{\mathbf{X}_{ij}^2}{p_{ij}} \mathbf{e}_i \mathbf{e}_i^T \right) - \mathbf{X} \mathbf{X}^T.
\end{aligned}$$

Notice that sampling according to the element-wise sampling probabilities of eqn. (1) satisfies  $p_{ij} \geq \frac{\beta}{2} \frac{\mathbf{X}_{ij}^2}{\|\mathbf{X}\|_F^2}$  and so we get

$$\mathbb{E}[\mathbf{M}_t \mathbf{M}_t^T] = \sum_{i,j=1}^{m,n} \left( \frac{\mathbf{X}_{ij}^2}{p_{ij}} \mathbf{e}_i \mathbf{e}_i^T \right) - \mathbf{X} \mathbf{X}^T \preceq \frac{2 \|\mathbf{X}\|_F^2}{\beta} \sum_{i,j=1}^{m,n} \mathbf{e}_i \mathbf{e}_i^T - \mathbf{X} \mathbf{X}^T = \frac{2n \|\mathbf{X}\|_F^2}{\beta} \mathbf{I}_m - \mathbf{X} \mathbf{X}^T.$$

Using Weyl's inequality we get

$$\|\mathbb{E}[\mathbf{M}_t \mathbf{M}_t^T]\|_2 \leq \max \left\{ \|\mathbf{X} \mathbf{X}^T\|_2^2, \frac{2n \|\mathbf{X}\|_F^2}{\beta} \|\mathbf{I}_m\|_2 \right\} = \frac{2n}{\beta} \|\mathbf{X}\|_F^2.$$

We can now apply Theorem 2 with  $\rho^2 = \frac{2n}{\beta} \|\mathbf{X}\|_F^2$  and  $\gamma = \frac{3\sqrt{mn}}{\beta} \|\mathbf{X}\|_F$  to conclude that  $\|\mathcal{S}_\Omega(\mathbf{X}) - \mathbf{X}\|_2 \leq \epsilon$  holds subject to a failure probability at most  $\diamond$

$$(m+n) \exp \left( \frac{-s\beta\epsilon^2}{4n \|\mathbf{X}\|_F^2 + 2\epsilon\sqrt{mn} \|\mathbf{X}\|_F} \right).$$

Setting the failure probability equal to  $\delta$ , we conclude that it suffices to set  $s$  as follows:

$$s \geq \frac{1}{\beta\epsilon^2} (4n \|\mathbf{X}\|_F^2 + 2\epsilon\sqrt{mn} \|\mathbf{X}\|_F) \ln \left( \frac{m+n}{\delta} \right).$$

We now consider two cases. First, if  $\epsilon \leq \|\mathbf{X}\|_F$ ,

$$\begin{aligned}
4n \|\mathbf{X}\|_F^2 + 2\epsilon\sqrt{mn} \|\mathbf{X}\|_F &\leq \max\{m, n\} (4 \|\mathbf{X}\|_F^2 + 2\epsilon \|\mathbf{X}\|_F) \\
&\leq 6 \max\{m, n\} \|\mathbf{X}\|_F^2,
\end{aligned}$$

which immediately proves the first case of Theorem 1. Similarly, if  $\epsilon > \|\mathbf{X}\|_F$ ,

$$4n \|\mathbf{X}\|_F^2 + 2\epsilon\sqrt{mn} \|\mathbf{X}\|_F \leq 6\epsilon \max\{m, n\} \|\mathbf{X}\|_F$$

and the second case of Theorem 1 follows.

## References

- [1] D. Achlioptas, Z. Karnin, and E. Liberty. Matrix entry-wise sampling: Simple is best. In *Neural Information Processing Systems*, 2013.
- [2] D. Achlioptas and F. McSherry. Fast computation of low rank matrix approximations. In *Proceedings of Symposium on the Theory of Computing*, pages 611–618, 2001.
- [3] D. Achlioptas and F. McSherry. Fast computation of low-rank matrix approximations. *Journal of the ACM*, page 54(2):9, 2007.
- [4] P. Drineas and A. Zouzias. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality. In *Information Processing Letters*, pages 385–389, 111(8), 2011.
- [5] B. Recht. A simpler approach to matrix completion. In *The Journal of Machine Learning Research*, pages 3413–3430, 12, 2011.